Managing Large Datasets for Atmospheric Research

Dr. Gao Chen
NASA Langley Research Center
Hampton, VA 23681
gao.chen@nasa.gov

Overview

Field Studies vs. Satellite Observations

- Satellite Observations:
  - ✓Pros: global and long-term coverage – suitable for climatological studies and model assessment
  - ✓Cons: limited species coverage, coarse spatial resolution
- Ground-based Field observations:
  - ✓Pros: direct measurement for multiple species – suitable for in-depth case studies and assessment of satellite observations and model simulations
  - ✓Cons: limited spatial coverage
- Airborne Field Observations:
  - ✓Pros: Comprehensive suite of direct observations with high spatial and temporal resolutions and good spatial coverage – suitable for probing atmospheric processes, assessing satellite observations, and model
  - ✓Cons: Limited temporal coverage

Overviews of in-situ Atmospheric Measurements: Aerosol Microphysical Properties

- Number density measurements:
  - CN (> 10 nm), UfCN (> 3 nm), and nonvolatile CN (> 10 nm)
- Number size distribution measurements:
  - SMPS/DMA: ~10 – 300 nm in **mobility** diameter
  - UHSAS: ~0.06 – 1 μm in **optical** diameter
  - OPC/LAS: ~0.1 – 8 μm in **optical** diameter
  - APS: ~0.6 – 20 μm in **aerodynamic** diameter

It is important to recognize that particle size measurement depends on the technique/instruments and the calibration material.

- CCN concentration measurements at a given instrument super-saturation(s)
- Ice Nuclei concentration measurement
- g(RH) measurements: hygroscopic size growth factor

Overviews of in-situ Atmospheric Measurements: Aerosol Composition

- Water soluble composition measurements:
  - PILS/IC: submicron K+, Na+, Mg+, Ca++, NH4+, Cl-, NO3-, and SO4=
  - PILS/TOC: submicron water soluble organic carbon
  - Filter/IC: K+, Na+, Mg+, Ca++, NH4+, Cl-, NO3-, and SO4=
- Non-refractory composition measurements:
  - AMS: submicron NH4+, Cl-, NO3-, SO4=, and OA
  - SP2: BC mass Concentration for particle between ~60 – 800 nm
- Single particle composition measurements:
  - PALMS: number fractions for particle chemical composition

Overviews of in-situ Atmospheric Measurements: Aerosol Optical Properties

- Scattering coefficient measurements
  - Nephelometer
- Absorption coefficient measurements
  - PSAP, TAP (Tricolor Absorption Photometer), and PAS
- Extinction coefficient measurements
  - CRD-AES/CRDS (Cavity Ring Down Aerosol Extinction Spectrometer), CAPS (Cavity Attenuated Phase Shift Spectrometer)
- Phase function measurements
  - PI Neph (Polarized Imaging Nephelometer)
- Hygroscopic growth factor for scattering or extinction measurements
  - Dry and humidified nephelometers

Necessary Information for Proper Use of Aerosol Measurements

- Inlet Size Cut (e.g. aircraft inlet size cut typically at < 5 µm – aerodynamic diameter)

Note the size cut is typically defined by 50% transmission efficiency.

- Instrument Size Range
- Instrument Sizing Technique:
  - SMPS/DMA: Mobility Size
  - UHSAS, OPCs, LAS: Optical Size
  - APS: Aerodynamic Size
  - AMS: Vacuum Aerodynamic Size

  The instrument perceived size is a function of particle properties: including composition, density, shape, and refractive index
- Instrument size calibration type (e.g. Polystyrene Latex Spheres vs. Ammonium Sulfate)

Necessary Information for Proper use of Aerosol Measurements (Cont.)

- Detection Limits and Data flags
- Data Reporting Conditions: STP vs. Ambient
- Instrument Relative Humidity – typically aerosols are dried in the sampling inlet and/or instrument
- Tagged variables: CCN and SS, f(RH) and instrument RHs, g(RH) and particle size and instrument RHs.
- Data processing algorithms and assumed parameterization
- Contact Instrument PI for questions

Understand Measurement Uncertainties

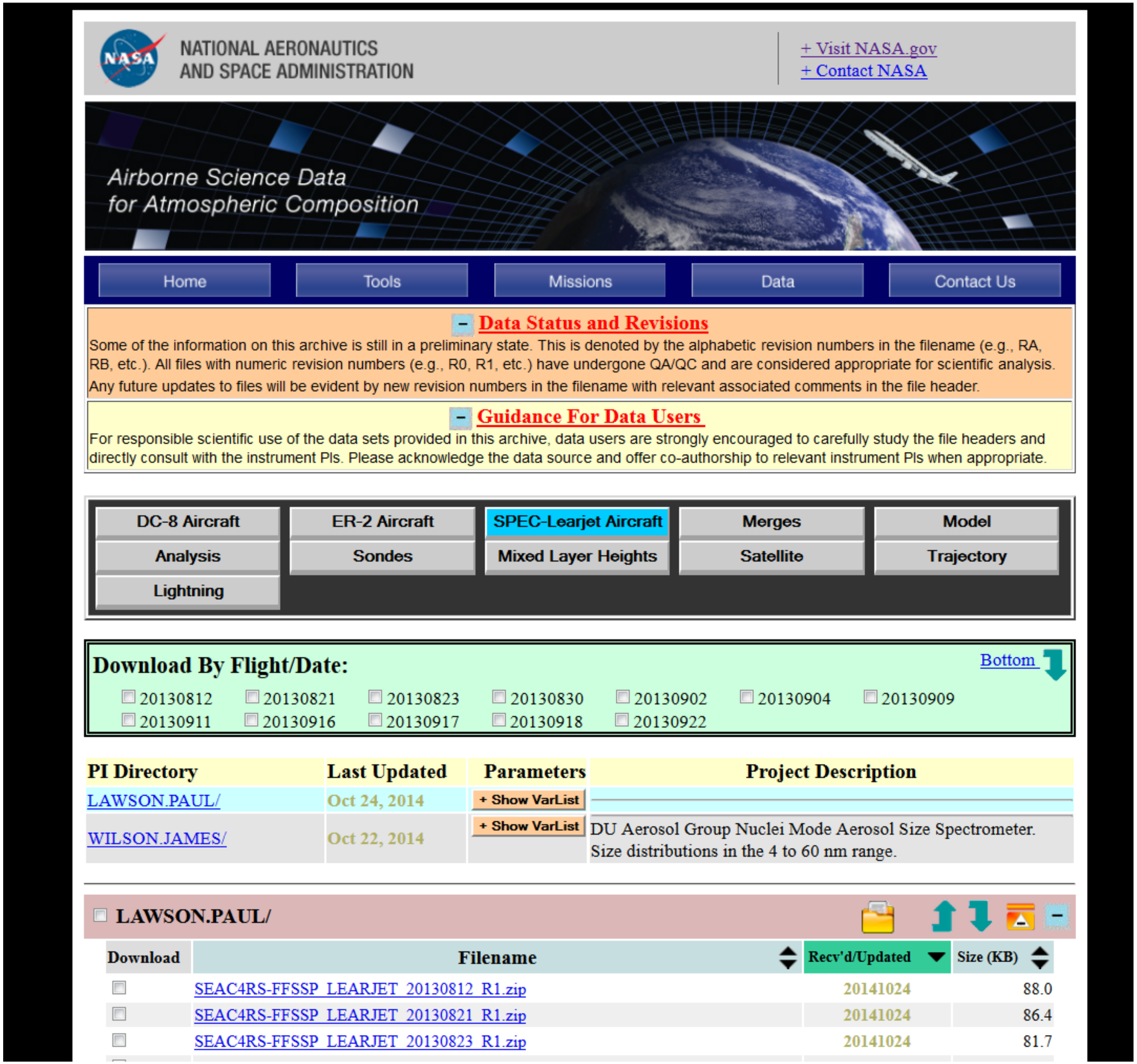
- Reported uncertainty as total, accuracy or precision?
- Size measurement/cut uncertainty
  - Type of size determination
  - Optical measurement dependence on particle refractive index and shape
- Difference in data processing algorithm and procedures
- Instrument/sampling timing uncertainties and timing difference between different instruments
- Measurement assumption, e.g., volume or surface area density estimates based on spherical particle assumptions.
- Difference between number based and mass based measurements
  - CCN concentration vs. AMS or PILS chemical composition
  - PALMS number fraction vs. Mass fraction

Project Data Websites for Airborne and Surface Measurements:

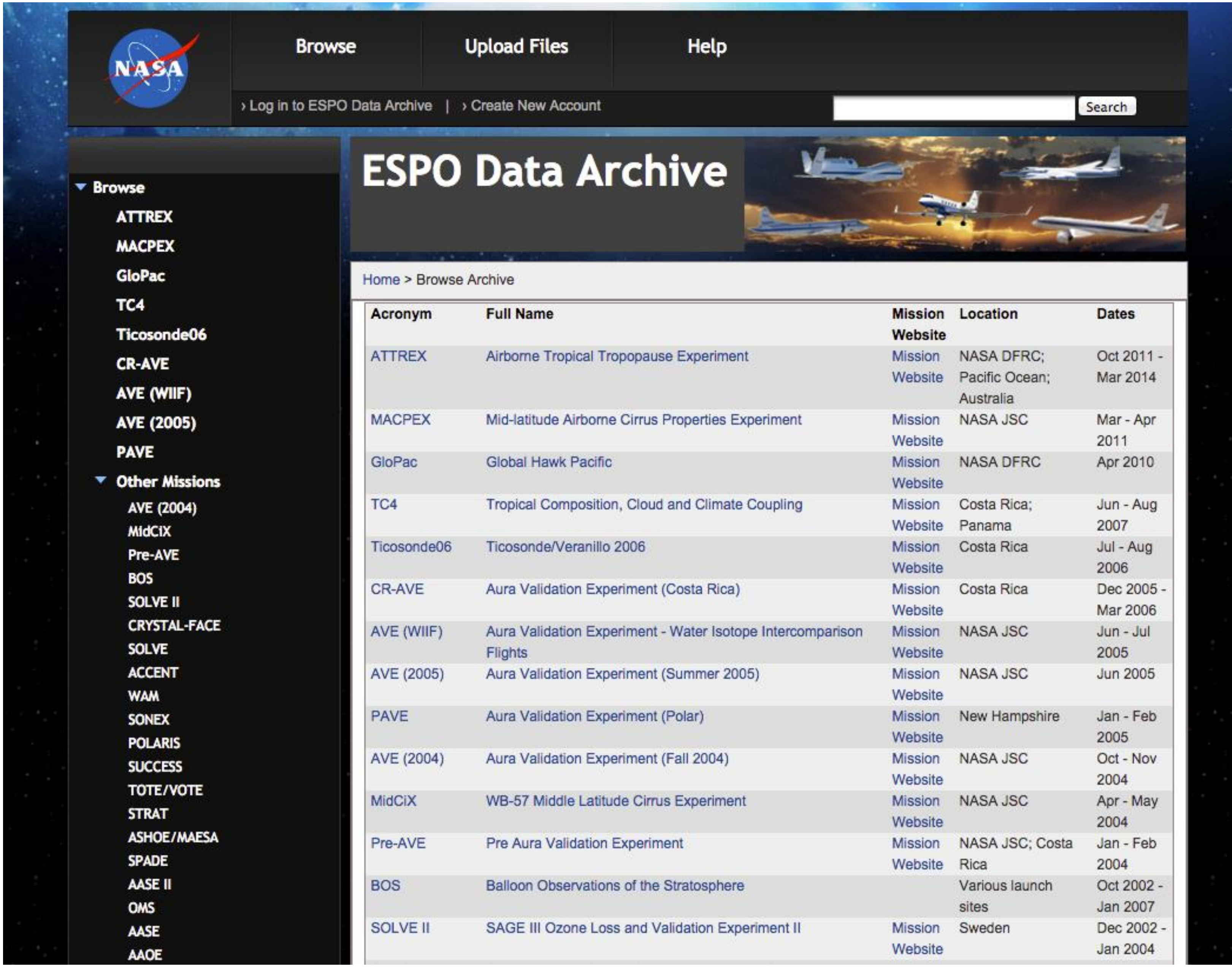
- http://www-air.larc.nasa.gov
  - Data holding: airborne field observations from NASA's tropospheric chemistry program sponsored field studies
  - Instrument PI data and merge data
  - Data organized by field projects, aircraft platform, and instrument PIs with some descriptions of measurement variables
  - Data files are typically in GTE or ICARTT format
- https://espoarchive.nasa.gov/
  - Data holding: airborne field observations from NASA's upper atmospheric research program sponsored field studies
  - Instrument PI data
  - Data organized by field projects and measurement categories with some descriptions of the measurement variables
  - Data files are typically in AMES or ICARTT format

Data Websites for Airborne and Surface Measurements (cont.):

- http://www.esrl.noaa.gov/csd/field.html
  - Data holding: airborne field observations from NOAA ESRL CSD field studies
  - Instrument PI data and merge data
  - Data organized by and field projects and instrument PIs with some descriptions of measurement variables
  - Data files are available in ICARTT format and IGOR format
- http://saga.pmel.noaa.gov/data/
  - Data holding: shipboard field observations from NOAA and NSF sponsored studies
  - Instrument PI data
  - Data organized by field projects and file names with some descriptions of the measurement variables
  - Data files are available in ICARTT, ACF, and IGOR formats



Available At: http://www-air.larc.nasa.gov



Acronym	Full Name	Mission Website	Location	Dates
ATTREX	Airborne Tropical Tropopause Experiment	Mission Website	NASA DFRC; Pacific Ocean; Australia	Oct 2011 - Mar 2014
MACPEX	Mid-latitude Airborne Cirrus Properties Experiment	Mission Website	NASA JSC	Mar - Apr 2011
GloPac	Global Hawk Pacific	Mission Website	NASA DFRC	Apr 2010
TC4	Tropical Composition, Cloud and Climate Coupling	Mission Website	Costa Rica; Panama	Jun - Aug 2007
Ticosonde06	Ticosonde/Veranillo 2006	Mission Website	Costa Rica	Jul - Aug 2006
CR-AVE	Aura Validation Experiment (Costa Rica)	Mission Website	Costa Rica	Dec 2005 - Mar 2006
AVE (WIIF)	Aura Validation Experiment - Water Isotope Intercomparison Flights	Mission Website	NASA JSC	Jun - Jul 2005
AVE (2005)	Aura Validation Experiment (Summer 2005)	Mission Website	NASA JSC	Jun 2005
PAVE	Aura Validation Experiment (Polar)	Mission Website	New Hampshire	Jan - Feb 2005
AVE (2004)	Aura Validation Experiment (Fall 2004)	Mission Website	NASA JSC	Oct - Nov 2004
MidCiX	WB-57 Middle Latitude Cirrus Experiment	Mission Website	NASA JSC	Apr - May 2004
Pre-AVE	Pre Aura Validation Experiment	Mission Website	NASA JSC; Costa Rica	Jan - Feb 2004
BOS	Balloon Observations of the Stratosphere		Various launch sites	Oct 2002 - Jan 2007
SOLVE II	SAGE III Ozone Loss and Validation Experiment II	Mission Website	Sweden	Dec 2002 - Jan 2004

Available At: https://espoarchive.nasa.gov/

Understanding ICARTT format Overview

- ICARTT file format is a NASA endorsed standard format for airborne field measurements and is widely used in airborne field studies sponsored by NASA, NOAA, NSF, and international partners, e.g. DLR and FAAM
- ICARTT format was first developed for the International Consortium for Atmospheric Research on Transport and Transformation field study in 2004. The NASA standards was established in 2010 and was revised in 2013
- ICARTT format is a comma delimited, ASCII-based, self-describing file structure, with a file header section for metadata and a data section. Detailed description can be found at: https://earthdata.nasa.gov/standards/icartt-file-format

Understanding ICARTT format: ICARTT File Example

File Structure Information and variable list

Required data description

Header Section

Data Section

DISCOVERAQ-LARGE-SP2_P3B_20130912_R0.ict

```
36, 1001
Anderson, Bruce E.
NASA Langley Research Center
Black Carbon Mass Concentration with an SP2
DISCOVER-AQ Texas
1, 1
2013, 09, 12, 2014, 01, 13
0
UTC_start, Seconds after midnight, Time of Aquisition
3
1, 1, 1
-9999.00, -9999.00, -9999.00
UTC_stop, Seconds after midnight, Time of Aquisition
UTC_mid, Seconds after midnight, Time of Aquisition
mBC, nm m-3, Black Carbon Mass Concentration
1
SPECIAL COMMENTS:BC mass-size distribution is available upon request.
18
PI_CONTACT_INFO: Bruce.E.Anderson@nasa.gov
PLATFORM: NASA P-3B Aircraft
LOCATION: The aircraft was stationed in Houston, Texas. Aircraft location
information in auxiliary files.
ASSOCIATED_DATA: N/A
INSTRUMENT_INFO: Droplet Measurement Technologies Single Particle Soot
Photometer (SP2)
DATA_INFO: Units: mBC, ng m-3 at 1013 hPa and 273.15 K
UNCERTAINTY: See PI; Nominally 20%
ULOD_FLAG: -7777
ULOD_VALUE: N/A
LLOD_FLAG: -8888
LLOD_VALUE: N/A
DM_CONTACT_INFO: Data Manager:Lee Thornhill;
email:Kenneth.L.Thornhill@nasa.gov
PROJECT_INFO: DISCOVER-AQ Texas
STIPULATIONS_ON_USE: Final Data.
OTHER_COMMENTS:       Data are reported at standard temperature and pressure
(1013 hPa and 273.15 K)
REVISION: R0
R0: Measurement timing has been synchronized with aerosol number
concentration.
UTC_start, UTC_stop, UTC_mid, mBC
50340, 50350, 50345, -9999.00
50350, 50360, 50355, -9999.00
50360, 50370, 50365, -9999.00
```

Understanding ICARTT format
File header description (I)

Number of header lines and format index

PI, affiliation, data description, and field project

File volume number, number of file volumes

Data collection date and Processing/Revision date

Data interval indicator

Independent Variable
Number of data variables
Scaling factors
Missing data flags
Data Variables

Special Comments

```
36, 1001
Anderson, Bruce E.
NASA Langley Research Center
Black Carbon Mass Concentration with an SP2
DISCOVER-AQ Texas
1, 1
2013, 09, 12, 2014, 01, 13
0
UTC_start, Seconds after midnight, Time of Acquisition
3
1, 1, 1
-9999.00, -9999.00, -9999.00
UTC_stop, Seconds after midnight, Time of Acquisition
UTC_mid, Seconds after midnight, Time of Acquisition
mBC, nm m-3, Black Carbon Mass Concentration
1
SPECIAL COMMENTS:BC mass-size distribution is available
upon request.
```

Understanding ICARTT format
File header description (II)

PI_CONTACT_INFO: Name, phone number, mailing address, email address.
PLATFORM: Platform or site information.
LOCATION: Including lat/lon/elev if applicable.
ASSOCIATED_DATA: File names for associated data: location data, aircraft parameters, ship data, etc.
INSTRUMENT_INFO: Instrument description, sampling technique and peculiarities, literature references, etc.
DATA_INFO: Units and other information regarding data manipulation.
UNCERTAINTY: Uncertainty information, if the uncertainty is not given as separate variables.
ULOD_FLAG: -7777 (Upper LOD flag, always -7's).
ULOD_VALUE: Upper LOD value (or function) corresponding to the -7777's flag in the data records.
LLOD_FLAG: -8888 (Lower LOD flag, always -8's).
LLOD_VALUE: Lower LOD value (or function) corresponding to the -8888's flag in the data records.
DM_CONTACT_INFO: Data Manager -- Name, affiliation, phone number, email address.
PROJECT_INFO: Study start & stop dates, web links, etc.
STIPULATIONS_ON_USE: (self explanatory).
OTHER_COMMENTS: Any other relevant information.
REVISION: Revision#, e.g., RA, RB, R0, R1...
REVISION Notes: Brief description of the data revision rationales.

Understanding ICARTT format: Data Flags used in data files

Missing Data Flag: -9s

Lower Limit of Detection Flag: -8s

Upper limit of Detection Flag: -7s

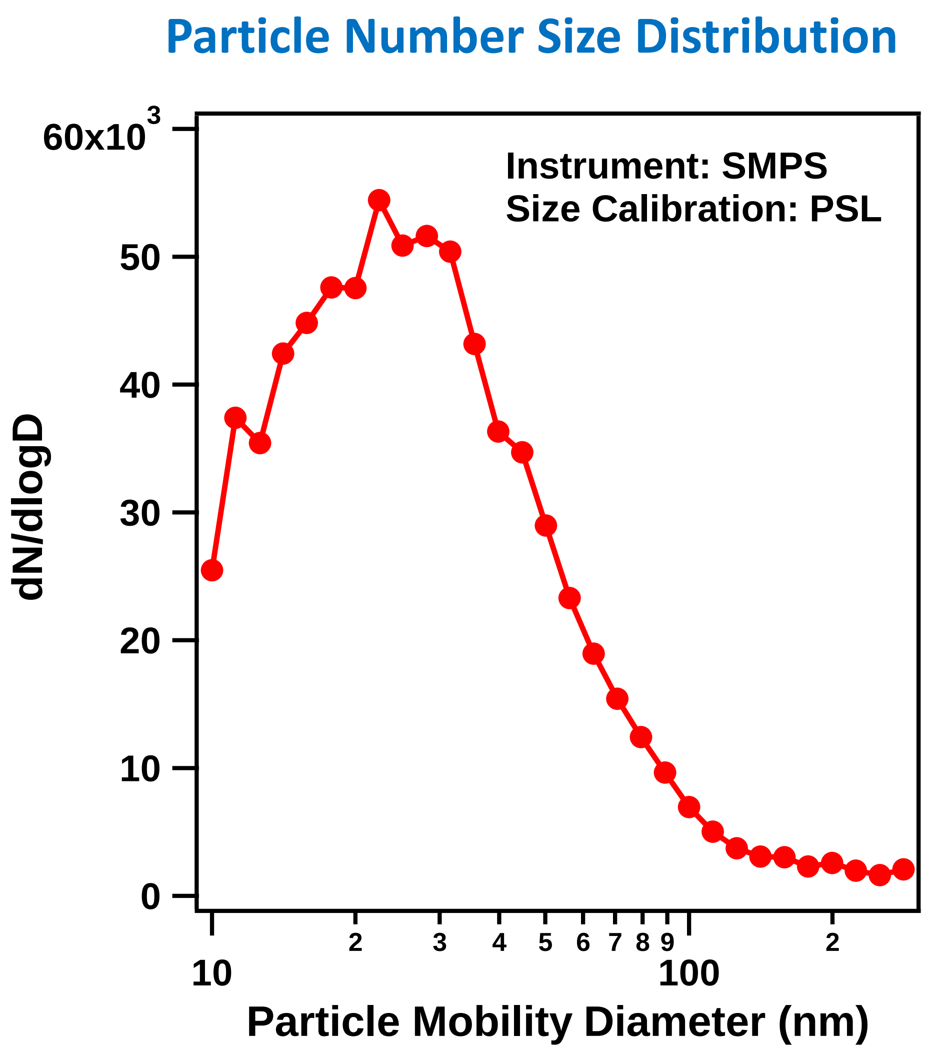
Data Quality Flag: Defined by measurement PI

Cloud Flag: Indicator if aircraft was sampling inside a cloud and a semi-quantitative measure of cloudiness

Sampling Location flag: Indicator of aircraft over specific sampling locations, e.g., over a ground site or in a spiral vertical profile

Data flags are typically not averaged

Understanding Particle Number Size Distribution Observational Data



Critical Information

- dN/dlogD or dN?
- Type of Instrument?
- Type Instrument Calibration?
- Bin definition – to increase S/N ratio, instrument bins may be combined by PI.
- Data reported at STP or ambient condition?
- Data Flags?
- Measurement Uncertainty
- PI Contact information
- Dry size distribution is common practice, actual instrument RH values may be obtained from the file header or by contacting PI

Understanding ICARTT format
Size Distribution Data Example

Sample File:

http://www-air.larc.nasa.gov/cgi-bin/ArcView/discover-aq.tx-2013, DISCOVERAQ-LARGE-SP2_P3B_20130912_R0.ict

Information from the File Header

SMPS_Bin06, #/cm3, Dry Aerosol Number Size Distribution (dNdlogD) by the SMPS (PSL mobility diameter)

INSTRUMENT_INFO: Size Distribution Data From the Scanning Mobility Particle Sizer (SMPS)

DATA_INFO: Aerosol measurements are reported at ambient temperature and pressure. Flagged data are due to missing data or to instrument issues

UNCERTAINTY: all data are 20%

OTHER_COMMENTS: SMPS data is rebinned to a standard 20 channels per decade rebinning scheme (dlogD = 0.05)

OTHER_COMMENTS: Mid-point diameters (in nm) are 10.0, 11.2, 12.6, 14.1, 15.8, 17.8, 20.0, 22.4, 25.1, 28.2, 31.6, 35.5, 39.8, 44.7, 50.1, 56.2, 63.1, 70.8, 79.4, 89.1, 100.0, 112.2, 125.9, 141.3, 158.5, 177.8, 199.5, 223.9, 251.2, 281.8

REVISION: R0

R0: Data is corrected over the version RA data for time lags

Similar types of information may be found in other size distribution data files

Understanding Particle Optical Property Observational Data

Optical measurements include scattering, absorption, extinction measurements, and hygroscopic dependence

- Measurement size cut due to inlet and instrument
- Data reporting conditions, i.e., STP vs. ambient
- Data reduction algorithm/procedures
  - Multiple reduction algorithms available PSAP
  - Anderson & Ogren correction for nephelometer is subject to larger uncertainties for supermicron particles
- Wavelength of the measurement
  - Different instruments make measurements at different wavelengths
  - Angstrom exponent wavelength conversion is an empirical relation
- Instrument RHs for f(RH) and g(RH) observation
  - f(RH) for hygroscopic growth is also based on an empirical relationship
  - g(RH) is size dependent

Understanding ICARTT format CRD-AES Extinction Data Example:

- Sample File:

http://www-air.larc.nasa.gov/cgi-bin/ArcView/seac4rs?DC8=1#WAGNER.NICHOLAS/, seac4rs-NOAA-AeroExt_DC8_20130906_R1.ict

- Important File Heard Information

INSTRUMENT_INFO: See Langridge et al., Aerosol Science and Technology 45:1305-1318, 2011.

DATA_INFO: Dry aerosol extinction in Mm-1 measured at RH < 30%, 273.15K and 101300Pa.

UNCERTAINTY: The precision varies with aerosol loading and is +/-20% at 10 Mm-1 and +/-5% at 100 Mm-1. The accuracy of the dry measurement is +/-2% and depends on RH for the humidified measurements, +/-5% at 85% RH and +/-15% at 90% RH. For more details, see Langridge et al., Aerosol Science and Technology 45:1305-1318, 2011

OTHER_COMMENTS: Measurements were made downstream of an impactor with the following pressure-dependent performance characteristics - 812 Pa: 90% transmission at 1141nm, 50% transmission at 1736nm, 10% transmission at 2376nm. 250 Pa: 90% transmission at 945nm, 50% transmission at 1388nm, 10% transmission at 2150nm.

REVISION: R1;R0

R1: In cloud data has been removed. Data permission altered.R0: Final Data. In cloud data has not been removed. Temporally synchronized with NOy. Data will be synchronized with DLH water when it is available.

Understanding ICARTT format PSAP Absorption Data Example:

- Sample File:

http://www-air.larc.nasa.gov/cgi-bin/ArcView/dc3#ANDERSON.BRUCE/, DC3-LARGE-ABS_DC8_20120601_R1.ict

- Important File Header Information:

INSTRUMENT_INFO: Absorption Data from Radiance Research 3-wavelength PSAPs
- The data were corrected for a variety of errors using Virkkula, AS&T, 44:706-712, 2010
- The new Virkkula correction is about 40% higher than the 2005 correction
- Flagged data are due to missing data or to instrument issues

DATA_INFO: Aerosol measurements are reported at standard temperature and pressure (0 deg C, 1013 mb)

UNCERTAINTY: all data are 20%

OTHER_COMMENTS: To convert from stdPT to ambient conditions, divide concentrations by stdPT

REVISION: R1

R1: adjusted time lags to matchup to NASA Langley's DLH

Understanding Particle Chemical Composition Observational Data

- Measurement size cut due to inlet and instrument
- Measurement techniques: water soluble vs. non-refractory
- Data reporting unit
- Measurement uncertainties and detection limits
- Measurement reporting conditions: ambient vs. STP. STP temperate definition may vary…
- Data quality statements and flags

Understanding ICARTT format Filter/IC Composition Data Example:

- Sample file:

http://www-air.larc.nasa.gov/cgi-bin/ArcView/dc3#DIBB.JACK/, DC3-SAGAAERO_DC8_20120530_R1.ict

- Important Header Information:

INSTRUMENT_INFO: Non-size selected ("bulk") aerosols are sampled isokinetically through forward facing aerosol inlet onto a teflon filter. Ions are extracted off of the filter into deionized water and ion chromatography analysis is performed.

DATA_INFO: Time data are reported in seconds after midnight UTC. Reported under STP (273 k & 1013 mb) CAUTION: Do not use these data on time scales shorter than those reported here.

UNCERTAINTY: It should be noted that the uncertainty of reported mixing ratios are generally about twice the listed LLOD value for each species. Data reported near LLOD should be treated with caution.

LLOD_FLAG: -8888
LLOD_VALUE:N/A,N/A,0.025,0.022,0.015,0.010,0.150,0.047,0.002,0.049,0.047,0.011

Sampling size cut information was NOT provided in the file header!

Understanding ICARTT format PILS/TOC Composition Data Example:

- Sample File:

http://www-air.larc.nasa.gov/cgi-bin/ArcView/discover-aq.ca-2013#ANDERSON.BRUCE/, DISCOVERAQ-LARGE-pilsTOC_P3B_20130128_R0.ict

- Important File Header Information:

INSTRUMENT_INFO: BMI Particle-Into-Liquid Sampler (PILS) and Siever Total Organic Carbon (TOC) Analyzer. Flagged data are due to missing data or to instrument issues

DATA_INFO: Aerosol measurements are corrected to and reported at ambient temperature and pressure

UNCERTAINTY: 20%

LLOD_FLAG: -8888,-8888,-8888

LLOD_VALUE: N/A,N/A,0.05

STIPULATIONS_ON_USE: This is final quality controlled data. For responsible scientific use of the data sets provided in this archive, data users are strongly encouraged to carefully study the file headers and directly consult with the instrument PIs. Please acknowledge the data source and offer co-authorship to relevant instrument PIs when appropriate.

OTHER_COMMENTS: WSOC measurements are solely the carbon mass (ugC m-3). To determine total water-soluble organic matter a multiplier must be used. Values for this multiplier range between 1.6 and 2.1 but for preliminary work a value of 1.8 is suggested. Please see Turpin (AS&T '01), Hand (JGR ’07) and Malm (AE ‘07) for more details.

REVISION: R0

R0: Final Data, corrected from field data for dilution associated with the PILS instrument

Sampling size cut information was NOT provided in the file header.

Understanding ICARTT format AMS Composition Data Example:

- Sample File:

http://www-air.larc.nasa.gov/cgi-bin/ArcView/seac4rs#JIMENEZ.JOSE/, SEAC4RS-AMS-60s_DC8_20130816_R1.ict

- Uncertainty reported with data, e.g.,
  - Nitrate_lt_1um_AMS_60s, Nitrate_lt_1um_AMS_prec_60s, Nitrate_lt_1um_AMS_DL_60s
  - The variable name included size, instrument, and integration time information

- Important Header Information:

INSTRUMENT_INFO: AMS_Type=V; Ionization_type=EI; AMS_SamplingTime: 1 s; AMS_SamplingCycle: Fast Mode, 6 s closed and 46 s open (+5 s PToF, reported under a different DataID) , 1 min synchronization clock; AMS_LensType=standard; AMS_LensPressure=1.40 Torr (actively controlled); AMS_FlowRate=90.7 sccm; AMS_PrsTempFlowCal: 273K, 1 bar; AMS_STPConversionFactor=StdtoVol_AMS; AMS_IEOverAB=7.5e-13; AMS_Calibrations: every ~3 days AN mass Calibration (400 nm particles) as well as two PSL Sizer calibrations (100-900 nm); AMS_RIENH4=3.78; AMS_RIESO4=1.2 (both based on in-field calibrations); AMS_CE: Calculated using the algorithm from Middlebrook et al, AS&T, 46, 258â€"271, 2012 using 1 min data; AMS_PMCut= ~1 um aerodynamic; AMS_InletRH: < 50%; Refer to http://cires.colorado.edu/jimenez-group/wiki/index.php?title=ToF-AMS_Main for further details.

Understanding ICARTT format AMS Composition Data Example:

DATA_INFO: Reported under STP (273 K & 1013 mb), use provided conversion factor to calculate under ambient temperature and pressure. Data was taken on a 1 min clock with 6 s background and 46 s ambient measurements, reported in 1 s increments. The remaining time was either instrument overhead or was spent taking particle size segregated data (reported in a separate data id). Unlike the aerosol species concentrations, all the variables detailing the chemical composition of the organic fraction were thresholded for the organic detection limit to avoid averaging artifacts; the thresholded data can be averaged without any need for mass weighting (lower thresholds are expected for final data). An IceFlag (based on AMS measurements of particulate water and Zinc) is included to identify periods when artifacts due to ice impaction on the inlet (not shared with the other aerosol instruments on the plane) were observed; in general, AMS data seems to be impacted to a lesser extent by these artifacts than other instruments, but extra caution should be exercised when using data from these periods. An organic nitrate fraction is included as an experimental product (1 min data due to the very noisy data), see Fry et al, ACP 13, 8585-8605, 2013 for details; data is not reported during elevated dust or seasalt episodes. O/C, H/C and OM/OC ratios are reported using the new calibration proposed by Canagaratna, Jimenez et al 2014 ACPD, which results in on average 28% higher O/C and 7% higher H/C ratios than using the Aiken et al ES&T 2008 method. The transmission of the AMS lens was lower than PM1 (comparable to Liu et AS&T, 2007), hence a time dependent correction has been implemented based on an experimental transmission curve (determined during the campaign using ammonium sulfate aerosol) and the aerosol volume data reported by the NASA LARGE group. The correction used is provided for reference purposes only as PM1InletCorr_AMS. AmmBalance_lt_1um_AMS is the ammonium balance for inorganic ions and serves as an aerosol acidity indicator. An ammonium balance of 1 represents a neutral aerosol phase, while 0 indicates a strongly acidic one.

Understanding ICARTT format AMS Composition Data Example (cont.):

UNCERTAINTY: Accuracy estimate (2sdev): Inorganics 34%, Organics 38%, dominated by uncertainty in the particle collection efficiency due to particle bounce. Precision error (1s) for each species is given as a separate variable. Detection Limits are given based on the method of Drewnick et al, Atmospheric Measurement Techniques, 2(1), 33-46, 2009 and were confirmed by comparison with periodic blanks.
ULOD_FLAG: -7777
ULOD_VALUE: N/A
LLOD_FLAG: -8888

LLOD_VALUE: -9999, -9999, -9999, -8888

OTHER_COMMENTS: Timebase corrected for Inlet Lag. Size cut is less than 1 um aerodynamic for this flight, species concentrations have been corrected based on an experimental transmission curve.
REVISION: R1,R0
R1: Final Data, Species corrected for transmission losses, improved chemical species ratios, new O/C and H/C based on Canagaratna, Jimenez et al 2014 ACPD (see Datainfo). R0:Preliminary Data.

Basic Analysis Techniques: Estimating Central Tendency and Spread

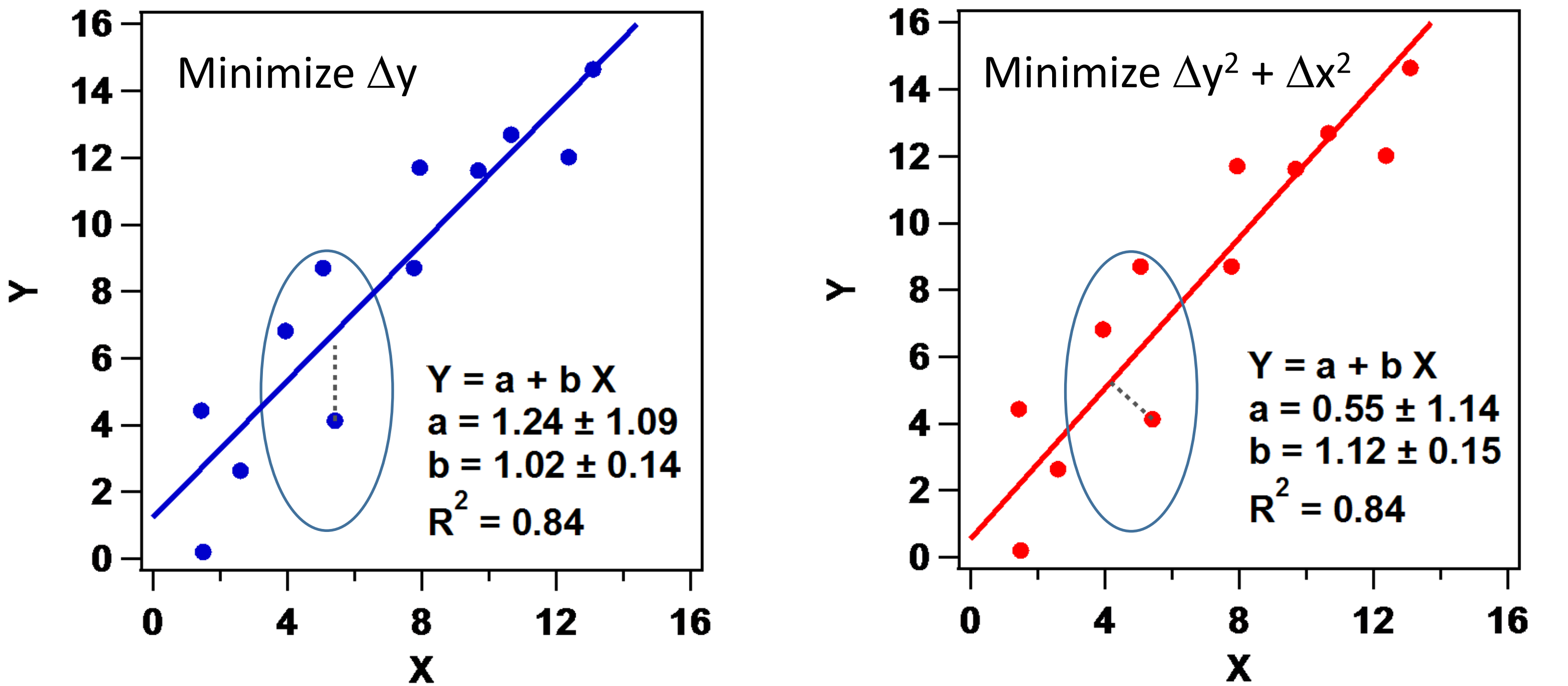
- The goal is to derive statistics that are robust and representative
  - Robustness: stats not sensitive to minor changes in sample size
  - Representativeness: stats accurately represent the data block
- Central tendency and Spread assessment

Data Distribution	Central Tendency	Spread
Normal	Mean, median, geometric mean	Standard deviation, percentiles, geometric standard deviation
Skewed, e.g., lognormal	Median, geometric mean	Percentiles, geometric standard deviation
Multi-mode	Median	Percentiles

Basic Analysis Techniques: Estimating Central Tendency and Spread (cont.)

- **Always check the data distribution first to see if the distribution is severely skewed**
  - ✓ Quick tests:
    - standard deviation to average ratio
    - comparison between median and mean
  - ✓ Normal distribution stats may be adequate if standard deviation is less than 1/3 of mean
  - ✓ If the distribution is "severely" skewed, median and geometric mean are preferred. The data may need to be divided into sub-groups to ensure robustness of the stats
  - ✓ Look for outliers and mitigate the influence of the outliers
- Median and percentiles can better handle LOD flags
- The LOD flags and values
  - LLODs represents upper limit values. LLOD/2 may or may not be the best estimate
  - Look for data trend to see if the LOD is an anomaly within a data block. If it is, one may consider if this an measurement issue
- Missing data flags should be removed before calculation for stats

Basic Analysis Techniques: Linear Regression



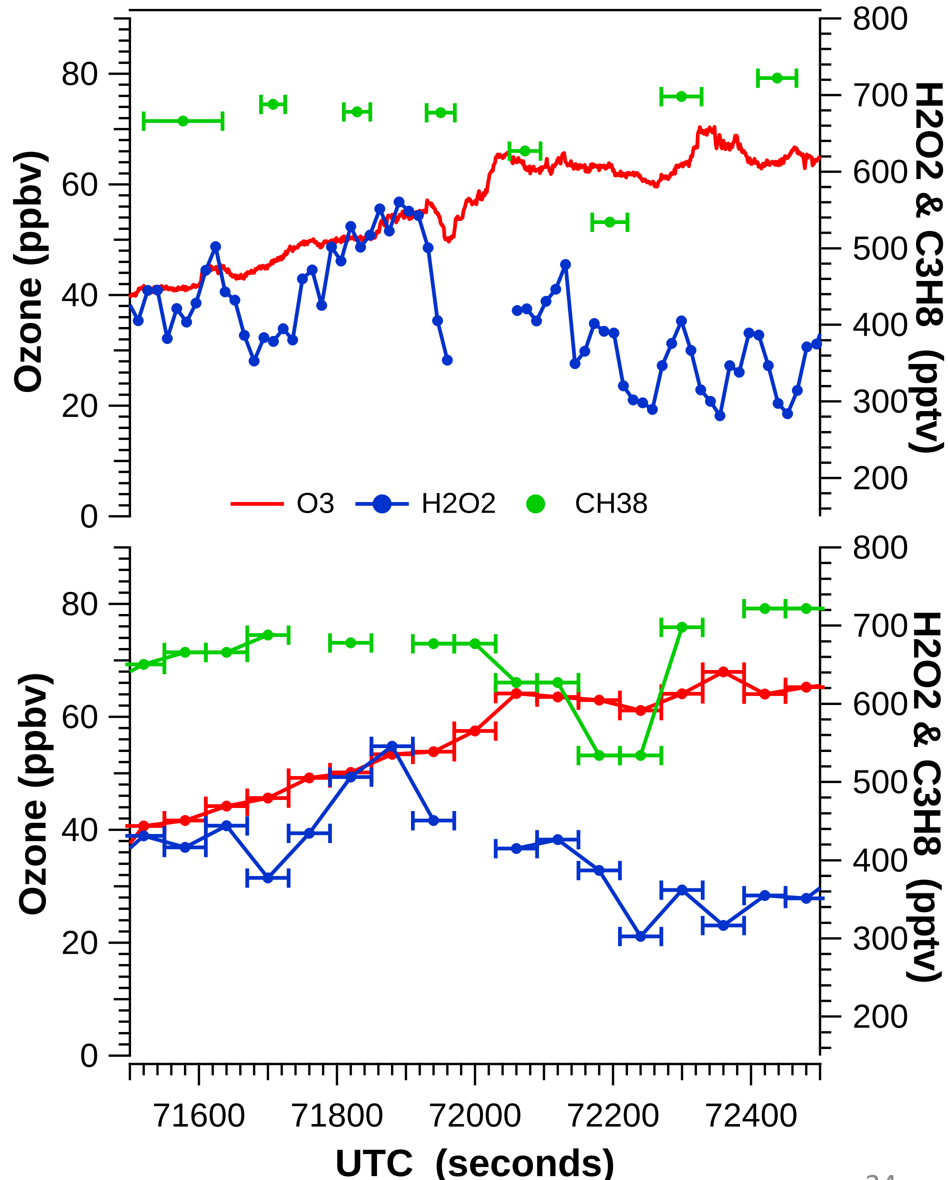
- Simple regression vs. Orthogonal regression
  - Use simple regression if X uncertainty is far smaller. ODR is not always better.
  - ODR regression gives consistent slopes for X vs. Y and Y vs. X, i.e., bb’ = 1, not R^2
- The range of X and Y can have substantial impact on R^2
- Not to use LODs in regression analysis
- Robustness test to minimize the impact of influential points/outliers

Basic Analysis Techniques: Data Quality Check

- Size distribution: integrated value should match the sum of the size bins
- Convert to common size when creating size distribution from multiple instruments
- Density check when independent volume and mass measurements are available
- Scattering and absorption closure with number size distributions with assumptions of refractive index and spherical particle shape.
- Absorption and BC correlation
- Measurement comparison for the common parameters to assess the consistency between instruments and to provide an independent check on measurement uncertainties

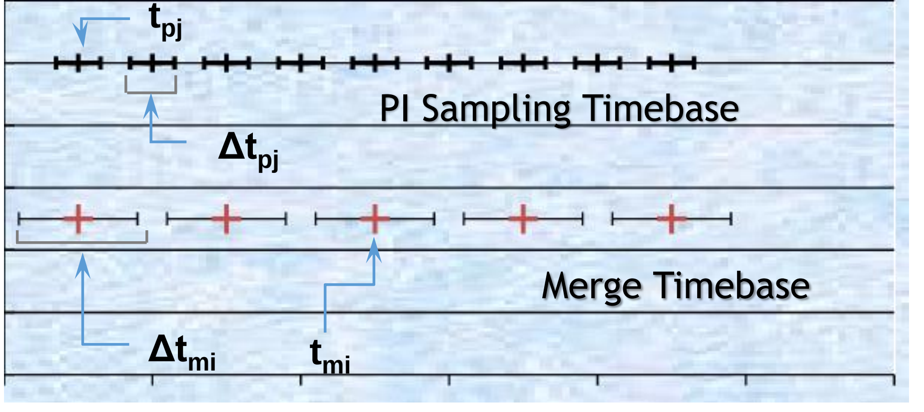
Data Merge:

- Data merge is an effective way to bring multiple measurement to a common time scale to examine the co-variation between difference atmospheric constituents
- Aircraft measurements are typically reported as a function of time. The merge process links measurements to sampling locations
- There are two kind of merges based on the time base: continuous time base and a specific measurement time base
- Merge files are often available at the airborne field study project data websites
- **Merge files typically do not contain full measurement description**



Weighted Merge Algorithm

The weighting factor $W_{i,j}$ is determined by the overlap between the j[th] PI measurement period and the i[th] merge time interval.



$$W_{i,j} = \begin{cases} \dfrac{(\Delta \mathrm{t_{mi}} + \Delta \mathrm{t_{pj}}) - 2(\mathrm{t_{mi}} - \mathrm{t_{pj}})}{\Delta \mathrm{t_{pj}}}, & if\ |\Delta \mathrm{t_{mi}} - \Delta \mathrm{t_{pj}}| < 2|\mathrm{t_{mi}} - \mathrm{t_{pj}}| < (\Delta \mathrm{t_{mi}} + \Delta \mathrm{t_{pj}}) \\ 0, & else\ if\ 2|\mathrm{t_{mi}} - \mathrm{t_{pj}}| \geq (\Delta \mathrm{t_{mi}} + \Delta \mathrm{t_{pj}}) \\ 1, & else\ if\ 2|\mathrm{t_{mi}} - \mathrm{t_{pj}}| \leq |\Delta \mathrm{t_{mi}} - \Delta \mathrm{t_{pj}}| \end{cases}$$

Scalar Variable Merge:

The average, x_i, and standard deviation, s_i, values for the i^{th} merge time interval are:

$$x_i = \frac{\sum_{j=1}^{n} w_{i,j}\, x_{pj}}{W_i} \quad \text{and} \quad S_i = \sqrt{\frac{\left(\sum W_{i,j}\, x_{pj}^2\right) - W_i x_i^2}{W_i \cdot W_i - \sum W_{i,j} \cdot W_{i,j}}} \quad \text{where} \quad W_i = \sum_{j=1}^{n} W_{i,j}$$

These formulas can be readily applied to a majority of the data variables, e.g., trace gas mixing ratios, aerosol composition concentrations, aerosol number densities, and aerosol optical properties (scattering coefficients, absorption coefficients, and extinction coefficients).

Having both average and standard deviation is necessary to examine the representativeness of the measurement.

Merging LOD and Uncertainties reported in Data Stream

- Background:
  - More PIs report uncertainties specifically for each measurement interval
  - Measurement precision typically increases with longer integration period
  - ULOD cannot be treated with reasonable assumptions
- Handling point-by-point uncertainties and LLODs:
  - If precision is reported: $\sigma_i^{(p)} = \frac{\sqrt{\sum W_{i,j}(U_{pj})^2}}{W_i}$ (quadrature average)
  - If accuracy is reported: $\sigma_i^{(a)} = \frac{\sum_{j=1}^{n} W_{i,j} U_{pj}}{W_i}$ (weighted average)
  - If total uncertainty is reported, they can be used as lower and upper limits for the uncertainty associated with the i^{th} merge time interval
- LLOD is driven by precision, thus: $LOD_i = \frac{\sqrt{\sum(W_{i,j} LOD_{pj})}}{W_i}$

Treating Data Flags in Merge

General Rule: treating the data flags based on research needs. The goal is to ensure the result is robust.

General Approach:

A. The merge interval is flagged if one point within the interval is flagged
B. The merge interval is flagged if certain percentage of the points within the interval are flagged; otherwise the flags are ignored

- LOD flags handling (when LOD values are not given for every data point):
  - Use approach A
  - Check the values not being flags to see if these values are significantly above (under) from the reported LLOD (ULOD) values. If they are, one needs to determine if this is due to ambient variation or instrument issues and should consult with the instrument PI
  - The field project merge files typically are generated using approach A
- Data Quality Flags:
  - Use approach A to ensure the highest data quality
  - Depending on the nature of the data quality issues, approach B should be considered to enhance the analysis sample size. Consulting with the instrument PI is highly recommended.

Treating data flags in Merge (cont.)

- Cloud Flags
  - Cloud flags are set according to the in-situ measurement of the condensed water concentration. This semi-quantitative flag is used to indicate the level of cloudiness: 0 for clear sky, 1 for thin cloud/uncertain, 2 for cloud, and 3 for heavy cloud
  - Average tends to work most of the time
- Sampling Location Flags
  - Use Approach B
  - Percentage threshold depends on data homogeneity and research objective

Special Case Merge

- Aircraft heading and wind direction need to use vector average for merge, i.e., use the x and y components. Detailed information can be found at: https://tad.larc.nasa.gov/media/TAD-Documentation.pdf in sections 3.3 and 3.4
- For ratios of variables, the ratios for merge intervals need to be recalculated after merging these variables individually. This is especially important when one or both variables has large variability.

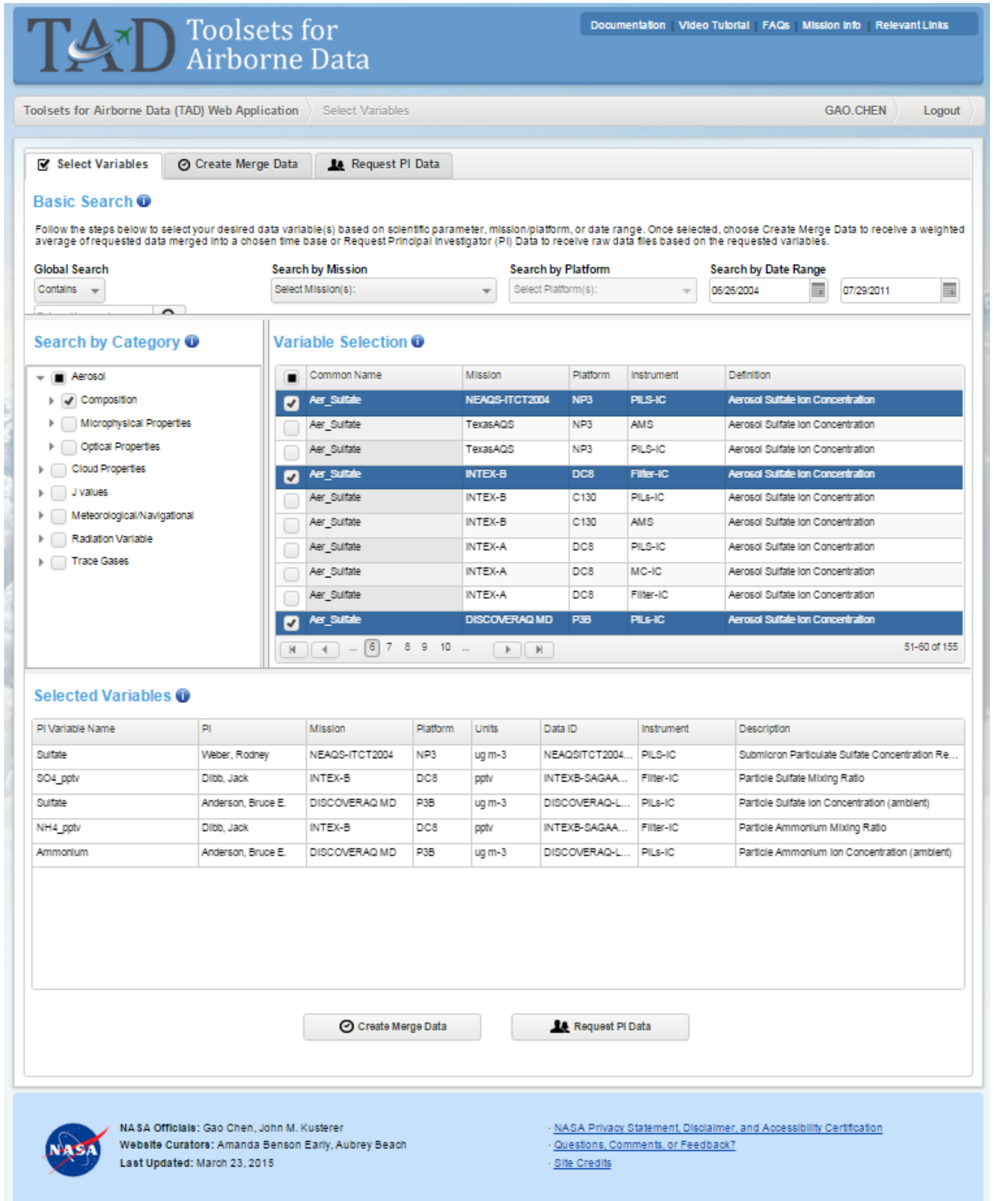
Merge Time Scale Selection

- Merge time scale selection should be determined by the nature of research
- For plume case analysis, one should use 1 sec. if 1 sec. data for key measurements are available
- For model assessment: the merge time scale should be chosen so that the aircraft sampling spatial range is compatible with the model grid size, i.e.

  Aircraft ground speed X merge time scale ≈ Model Grid Size

  - Know the temporal and spatial scale mismatch issue
  - Look for ways to evaluate the homogeneity at the subgrid scale. Difference between model and observations may be interpreted both correct
  - Check number of individual measurements within a model grid
  - Check data variability, e.g. standard deviation
  - In past studies, modelers have used 60, 300, 600, and 900 seconds
- If the preferred merge data is not available, one may try to make a request through the project data websites or use TAD

Toolsets for Airborne Data (TAD): Overview



- TAD is designed to be a central web portal to easily filter and download in-situ airborne data
- TAD enables data search across airborne field studies through text-based global search or variable categories and lists
- TAD supports access to PI data files
- TAD has tools to perform customized data merges
- TAD data visualization tools are currently under-development
- TAD is open to public, but registration (taking less than 3 min.) is needed for access.

TAD: Current and Pending Data Sets

Airborne Study	Deployment Year	Funding Agency	Ingest Status
DISCOVER-AQ MD	2011	NASA	Done
INTEX-A	2004	NASA	Done
INTEX-B	2006	NASA	Done
ARCTAS	2008	NASA	Done
NEAQS-ITCT 2004	2004	NOAA	Done
DISCOVER-AQ CA	2012	NASA	On-going
TexAQS 2006	2006	NOAA	Done
ARCPAC	2008	NOAA	Done
CalNEX	2010	NOAA	Done
DC3	2012	NASA & NSF	On-going
DISCOVER-AQ TX	2013	NASA	Pending
SEAC4RS	2013	NASA	Pending
DISCOVER-AQ CO	2014	NASA	Pending

TAD Merge Features:

- Option to choose variables of interest

  Latitude, longitude, pressure altitude, ambient pressure, and static temperature are default variables.
- Continuous time base from 1 to 3600 sec. or selected variable measurement time scale.
- Provide descriptive statistics, i.e., average, standard deviation, and number of points for each data merge time intervals
- Option to process uncertainties and LOD reported as a data stream.
- Option to select from one flight to all flights of a given mission and aircraft platform

Closing Remarks

- Understand the data – please read the PI file header carefully
- Know measurement assumptions
- Accurately represent the data
- **Contact PI for questions and advice**